面向叙词表更新的新术语分布特征研究*

■ 雷晓　常春　刘伟

中国科学技术信息研究所　北京 100038

摘要：［目的/意义］为增强叙词表实用性，需要不断地将领域中出现的新术语更新到叙词表中，更新维护过程中，从时间及词频等角度对新术语分布特征进行探索研究，可以为新术语发现方法提供参考。［方法/过程］基于新术语相关特征，结合对应文档频率在时间点和时间段上的发展分布，通过相关统计分析，研究术语在不同成长时期的分布特征，尤其界定术语在开始期与成长期的分布差异。［结果/结论］实证分析表明新术语一般处于术语发展的成长期，当候选新术语保持正向增长趋势超过一定年限，可以认为该术语同时具有新颖性、时间持续性及术语性特征。基于该分布特征进行领域新术语的识别，结合词表编制专家的判断，该方法在新术语收录判断中具有较高的准确率，且能有效识别实际应用中占比较多的低频词。

关键词：叙词表更新　新术语　时间分布　文档词频分布

分类号：G254.2

DOI：10.13266/j.issn.0252-3116.2019.20.014

叙词表又称主题词表，是以规范化、受控的和动态性的术语作为基本成分，用于标引、存储和检索文献的一种情报语言词汇表[1]。如今，叙词表不仅是图书情报学科重要的基础工具[2]，更被广泛应用于自然语言处理、机器翻译、知识抽取、数据挖掘、本体构建等研究领域。在信息海量激增的当下，优质且更新维护及时的叙词表或术语集合，是以上各研究领域的重要基石。随着很多学科领域不断深入快速发展，有很多已出现且需要及时被该领域叙词表收录的术语，本文称这部分术语为新术语。及时发现新术语以更新相应的叙词表，对于把握学科领域发展及相关研究应用都具有重要作用[3]。当前，叙词表的编制和更新主要依赖于专家的手工劳动，而网络环境下各个专业领域中新术语大量涌现，使得手工更新的方式远远滞后于新术语的增长速度，需要研究科学的方法发现新术语以提高叙词表更新效率，使之紧跟领域的发展。本文以《汉语主题词表》[4]中已收录术语为例，获取文献数据库中对应领域的文献元数据，根据文献关键词在对应发表年份的文档词频值形成关键词对应的时间序列，基于新术语的新颖性、时间持续性以及术语性等特征，利用统计分析方法研究术语文档频率（document frequency，DF）随时间的成长分布，探索术语在成长不同时期的分布差异，尤其界定术语从出现到成熟阶段的相关特征，以探究候选新术语满足什么条件可以更新到叙词表中，最后结合该术语分布特征识别领域新术语。

1　相关研究

新术语发现是叙词表更新维护的重要一环，相关研究主要集中在候选术语的自动获取，以及对其进行识别判断上，即判断候选新术语的成熟程度，确保叙词表收录的新术语不是偶发词，出现一段时间就不再被使用，而是兼具新术语的新颖性、时间持续性、专业性及规范性相关特征。

1.1　候选术语获取

叙词表更新维护中，首先需要获取尽可能多具有独立成词能力的词作为候选术语，再利用叙词表已收录术语及停用词等进行过滤获得候选新术语。目前，叙词表选词来源一般包括：文献提供的关键词、用户使用的检索词、各类词典中的专业词汇资源、用户通过相关平台给出的推荐词以及网络词汇等；也包括从自然

* 本文系国家社会科学基金项目“面向知识组织系统的新术语抽取研究”（项目编号：16BTQ087）和 国家科技图书文献中心“下一代国家科技创新知识服务开放系统”先期研发任务课题“STKOS 超级科技词表内容建设机制与发展研究（工学部分）”（项目编号：XQYF0101-2）研究成果之一。

作者简介：雷晓（ORCID：0000-0001-9984-2686），硕士研究生；常春（ORCID：0000-0003-2829-2589），研究馆员，博士；刘伟（ORCID：0000-0003-2857-5474），副研究员，博士，通讯作者，E-mail：liuw@istic.ac.cn。

收稿日期：2018-12-03　修回日期：2019-05-15　本文起止页码：121-128　本文责任编辑：王传清

语言中抽取获得的候选新术语。

通过自然语言处理获取得到候选新术语的方法主要有基于规则、基于统计和将两者融合的方法。基于规则的术语发现方法指的是利用文本语料的语言结构特征,制定例如词性规则、组块规则、停用词过滤规则、字串长度过滤规则等识别术语;还有一种就是基于词的上下文边界特征来制定规则[5],比如后停词过滤规则、前缀过滤词典、后缀过滤词典、相邻词过滤规则等。基于统计的术语发现方法,一种是基于传统的词频、互信息、最大似然比、TF-IDF 等统计方法;另外一种则是基于机器学习的方法,主要集中在有监督的机器学习方法上[6],例如支持向量机、隐马尔科夫模型、最大熵模型、条件随机场、规则学习算法、朴素贝叶斯、N-Gram[7]等。无论基于规则还是基于统计都各有利弊:基于规则的方法比较繁琐且通用性差,基于统计的方法噪声结果比较多[8]。目前,主要采用的是将两者融合的混合策略:一是采用规则先获取候选词汇,再通过统计的方法得到最终结果;二是先统计、再通过规则来识别要发现的术语;三是同时有效融合语言学和统计学特征的方法。比较有代表性的混合策略方法有 K. Frantzi[9]提出的 C-value/NC-value,以及在其基础上,由熊李艳等[10]、胡阿沛等[11]、韩红旗等[12]提出的改进策略,混合策略的使用有效结合了基于规则及基于统计方法的优势,可以识别一些多词术语、长术语以及嵌套术语等低频术语,效率较高。

1.2 新词识别判断

新术语发现一般需要在候选新术语的基础上,依据新术语的相关特征在候选新术语中做进一步筛选,以判断该候选新术语是否可以增加到对应的叙词表中。“新”则必然考虑新术语出现的时效性及新颖性特征,M. Wang 等认为时间变化特征可以提供很多文本处理信息,特别是对于新词检测问题,他们利用与时间相关的动态特征构建了新词识别模型[13];黄轩等[14]、邹纲等[15]、吴悦等[16]在各自的新词发现研究中都以某一时间为界,将语料分为背景语料和前景语料,认为如果某一候选词在背景语料中很少,而在前景语料中大量出现,则它很可能是一个新词。识别判断新词的方法通常可应用到新术语识别中,但其只考虑到了候选术语出现的时间节点问题,认为只要一个词自某时间节点出现,且词频达到一定程度则可以认为其满足被收入词表的条件。事实上,有大部分新词在出现后很快就消亡了,只有少部分新词能存活下去,继而逐渐发展为术语[17]。

总之,就目前的研究现状而言,抽取获得候选新术语的方法相对较多且比较成熟,但候选新术语一般不能直接作为新术语增加到叙词表中,需要进行一定的识别判断,目前该工作主要依靠领域专家进行人工判断,不符合大数据时代新术语量激增的现状,但相关研究依然较少且存在一定局限性。例如,人工判定新术语具有主观性,且效率低不及时[18];新术语形成时间具有模糊性,新词汇出现不代表会持续在领域内通用成为术语;术语本身由于不同专指程度在实际应用中会有数据量的较大差异,单从频数角度可能会把累计权重不高的术语直接剔除等。因此有必要根据新术语相关特征研究新术语的成长分布情况,特别是界定术语从出现到成熟阶段的分布变化,以探究候选新术语满足什么分布特性可以更新到叙词表中,而不单从出现时间及词频数量的角度进行判断。

2 新术语相关特征及成长分布特征

所谓新术语是指未曾收录到相应领域词表中,且在领域内某一时间节点之前没有出现过,或虽然偶发但曾经没有持续出现时间的术语。总结新术语相关特征,基于其新颖性、时间持续性以及术语性特征研究新术语的时间分布及文档词频分布特征,并以《汉语主题词表》中已收录术语为例,探究不同词频量水平、不同生命周期长度术语的成长分布状况。

2.1 新术语相关特征

判断一个候选新术语是否可以更新到叙词表中,其应同时具备两方面的特征:一是具有术语的单元性、术语性特征。单元性是指术语必须具有独立成词能力;术语性则指术语要同时具有规范性和专业性两方面的特征。其中,规范性指的是术语在某一特定专业范围内被广泛使用,专业性指术语具有领域相关性,一般在特定领域中流通使用。二是具有新颖性、时间持续性特征。新颖性指从时间参照角度,新术语是自某一时间点以来首次出现的具有新词形、新词义或者新用法的词汇[19];从词表参照角度,新术语是指通过各种途径产生的具有目前词表中基本词汇所没有的新形式、新意义或新用法的词语[20]。时间持续性特征则指该术语的存在是持续的,而不是在出现后很快就会消亡。本次研究重点关注的是具有新词形的新术语,因此,下文中提到的新术语均指此类。

依据以上新术语相关特征,并假设候选新术语已获取,即候选新术语已经确定是领域内未收录的、具有独立成词能力的词。本文以《汉语主题词表》中已收

录术语的成长分布为例,探究关于判断候选新术语能否加入到词表中的相关分布特征方法。

2.2 基于新颖性及时间持续性的新术语时间分布特征

新术语具有新颖性,同时也必须具有时间持续性。利用词表已收录术语对候选术语进行过滤,获得未出现在叙词表中的候选新术语,对候选新术语,在时间分布上有如下特性:自某个时间点起的一定时间段内候选新术语持续出现,其中时间点是指具体的某个时间戳,具有定位性,例如某一年甚至某一个具体日期;而时间段则指选定的一个时间范围,具有历程性,是有起点、有终点还有长度范围的一段时间[21]。

用三元组(w, t_i, df)表示候选新术语 w 在第 t_i 年的文档词频为 df,记在统计时,w 的最早时间点为 t_0,当 t_0 同时也是该术语一个连续出现时间段 T 的起点,则称 t_0 为候选新术语的出现时间点 s,其中时间段 T 越长越好,在新术语识别研究中,一般选取离研究较近的时间点作为时间段的终点。若 t_0 只是一个偶发值,即 $df(t_1)=0$,则选择往后新的出现时间点作为 t_0 进行判断,直至找到 s。例如通过中国知网数据库(http://www.cnki.net/)统计电力工业领域内含“无线充电”一词的文章数,即文档词频,统计结果见表 1。术语“无线充电”在文献库中最早出现的时间点为 2000 年,记 $t_0=2000$,而 $df(2001)=0$,$df(2002)=1$;则 $t_0=2002$,又 $df(2003)=df(2004)=df(2005)=0$,$df(2006)=2$;则 $t_0=2006$,$df(2006+i, i \leqq T)>0$,则“无线充电”一词的出现时间点 $s=t_0=2006$,则研究时间段定为 2006－2017 年。

表 1 “无线充电”一词文档词频逐年统计

年份 t_i	文档词频 df	年份 t_i	文档词频 df	年份 t_i	文档词频 df
2000	2	2009	3	2014	132
2002	1	2010	12	2015	169
2006	2	2011	19	2016	156
2007	4	2012	51	2017	198
2008	1	2013	88		

基于新颖性和时间持续性特征,提出结合时间点、时间段两个角度,判断术语出现时间点的方法,即选择一个连续时间段的起点作为观察术语成长分布的起点,而不是直接选择统计的最早出现时间作为术语成长分布的起点。事实上,当一个词汇在以年为时间点的研究时间段内出现断点,恰恰说明其不具有规范性,在领域内还尚未被广泛使用,无法称之为术语。

2.3 基于术语性的新术语文档词频分布特征

术语性包括规范性和专业性两个特征,规范性是指术语在某一特定专业范围内被广泛使用,专业性指术语具有领域相关性。再结合术语新颖性、时间持续性特征,即当候选新术语能被相关领域学者普遍接受,能在出现后较长一段时间,仍持续被使用在领域内不同的科技论文中时,可认为其是新术语。本文通过统计数据库中,某一领域历年包含某候选术语文献的文献频数,即文档频率的分布情况来反映术语的规范性与专业性。对分布情况进行研究,其意义在于不同术语的研究热度有所区别,相应的数据量差异较大,通用的单从频数角度判断术语通用性的方法存在其局限性。常春等[22]基于生态学理论“Logistic 生物种群增长模型”,将单个生物种群个体数量增长过程与叙词表术语对应文档频数增长过程相类比,总结术语词频变化规律,并提出术语成熟的生命周期变化特征,即根据对应的文档词频数量变化,将一个成熟术语的生命周期划分为开始期、成长期以及饱和期等多个阶段。

基于上述考虑,本文同样从文档词频分布角度出发,并假设存在一个特殊时间点,术语在时间点前属于“开始期”,词频少且无明显增长;而在该时间点后词频忽然增大,且在往后一定时间范围内保持正向增长趋势,称该时间点为候选新术语的“正向增长转折点”,而该时间点往后保持正向增长的时间段则称为术语的“成长期”;再往后术语增长幅度开始出现正负范围的小幅度浮动,但大体上保持在一个稳定的词频数量级上,称这一发展时间段为术语的“饱和期”。为有效量化分布趋势的波动情况,本文计算了时间段内术语文档词频逐年的环比增长率(ring growth, rg),见公式 1。当环比增长率值为正,表明术语文档词频相比去年有所增加,反之亦然。

词频环比增长率 rg = (本时间点词频 － 上一时间点词频)/上一时间点词频 　　公式 1

关于“正向增长转折点”的确定,用三元组(w, t_i, rg)表示候选新术语 w 在第 t_i 年对应的词频环比增长率为 rg,而 $rg=[df(t_i)-df(t_{i-1})]/df(t_{i-1})$,逐年计算 w 的 rg 值并判断其大小。假设第 t_i 年 rg 值为正,且 t_i 年往后超过 5 年以上 rg 值均大于阈值零,则 t_i 为该术语的正向增长转折点。若 $rg(t_i)$ 虽然为正值,但 $rg(t_{i+1})<0$,则选择往后新的 rg 值进行判断,直至找到新的正向增长转折点。例如根据表 1 中术语“无线充电”的文档词频统计其对应环比增长率,统计结果见表 2,虽然 $rg(2007)=1>0$,但 $rg(2008)=-0.75<0$,而 $rg(2009)>0$,且 2009 年往后超过 5 年的数据其对应 rg 值均大于 0,即该术语的正向增长转折点确定为 2009 年,

图书情报工作
第 63 卷 第 20 期 2019 年 10 月

一般当术语处于成长期超过一定年限,例如超过 5 年,即“无线充电”一词于 2014 年就可以考虑将其更新到叙词表中。另外,以上虽然以 0 作为对 rg 值判断的阈值,但应该允许有小幅度的误差存在,可根据实际情况调整该阈值大小,例如一个比较接近 0 的负数。

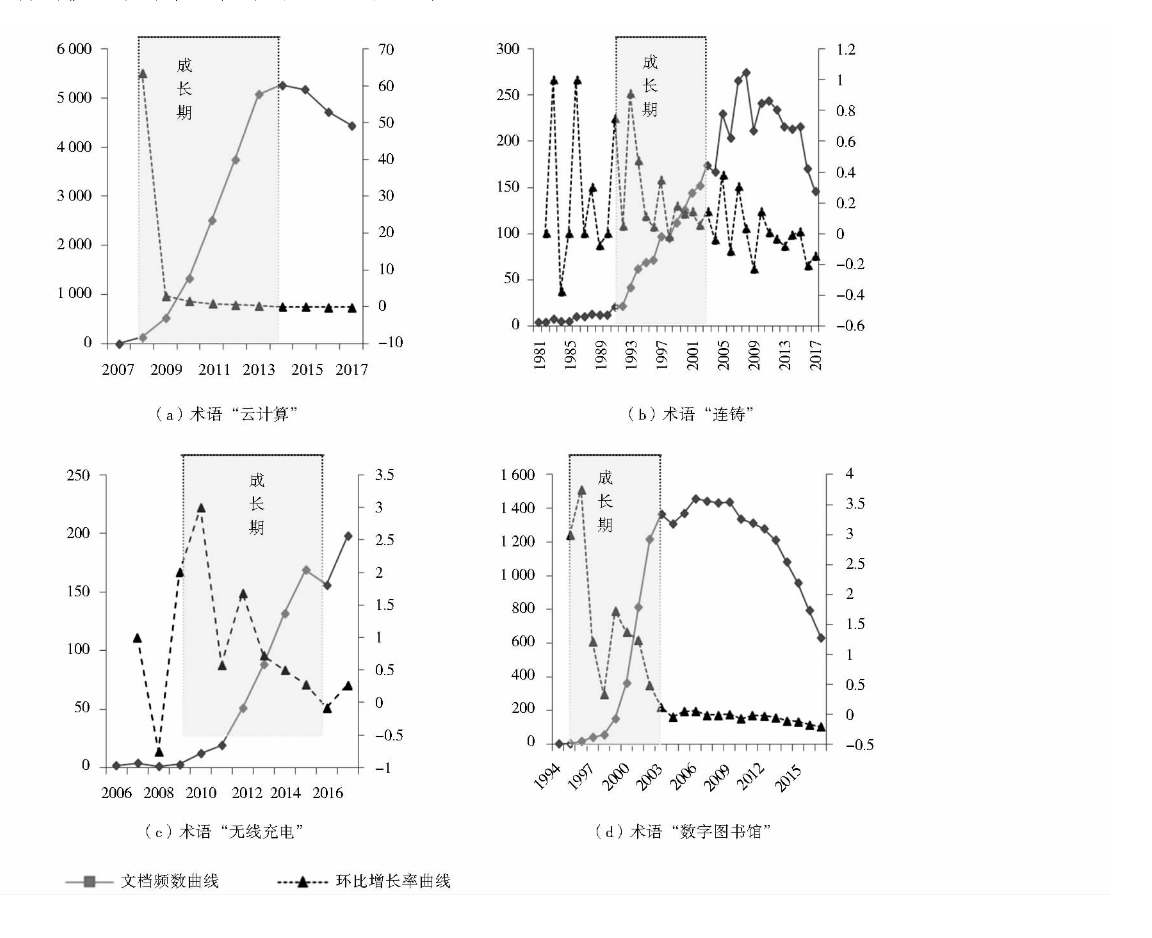
表 2 “无线充电”一词文档词频环比增长率(rg)逐年统计

年份 t_i	$rg(t_i)$	年份 t_i	$rg(t_i)$	年份 t_i	$rg(t_i)$
2006		2010	3	2014	0.5
2007	1	2011	0.583	2015	0.280
2008	-0.75	2012	1.684	2016	-0.077
2009	2	2013	0.725	2017	0.269

面向叙词表更新展开的新术语分布特征研究,重点区分术语开始期与成长期的分布差异。当候选新术语处于开始期,其对应文档词频量少且增长不明显,不认为其具有术语特征,尚不能更新到对应叙词表中。开始期的确定本身需要通过一段时间观察获得,为了及时将新术语更新到叙词表中,观察时间也不宜太久,需要确定相关阈值,例如确定正向增长转折点后的第 5 年及以上,可认为该术语可以更新到叙词表中。

2.4 4 个不同类别术语的分布特征

术语本身由于不同专指程度在实际应用中会有使用频次的较大差异,苏其龙[17]在新词发现研究中根据新词频数及生命周期长短,将新词分为以下 4 个类别:短期高频词、短期低频词、长期低频词和长期高频词。本文根据术语的文档词频统计数据,从连续出现时间段 T 的长度以及文档词频数量等级出发,当 T 为 15 年及以下,认为其为短期,大于 15 年则为长期;当最大 DF 值 <500,认为其属于低频词,反之为高频词,以此标准将 2014 版《汉语主题词表》中已收录术语同样分为以上 4 类,并依次以短期高频词“云计算”、长期低频词“连铸”、短期低频词“无线充电”、长期高频词“数字图书馆”为例,给出各自的文档分布统计情况及对应的环比增长率值,生成 4 个统计趋势图对比统一显示为如图 1 所示:



注:每个折线图左侧的主要纵坐标轴表示文档频数刻度,右侧次要纵坐标轴表示环比增长率刻度,横坐标轴表示年份;阴影部分是本文确定的该术语“成长期”阶段

图 1 4 个术语文档词频分布及对应环比增长率

雷晓，常春，刘伟. 面向叙词表更新的新术语分布特征研究[J]. 图书情报工作,2019,63(20):121－128.

图1中(a)是信息科技领域术语“云计算”一词的文档词频在近年来随时间的分布情况，以及对应的环比增长率结果，从时间点及时间段两个层次上，“云计算”一词的最早出现时间点 s 为2007年，逐年计算其历年环比增长率，发现 $rg(2008) = (129 - 2)/2 = 63.5$，而2008－2014年度，相关文献量一直保持增长趋势直至到达一个较高数据量水平上，称这一段时间为候选新术语的成长期；而自2015年往后，环比增长率出现正负范围较小幅度的波动，称这一时期为术语的饱和期。区别于成长期的正向增长趋势，这一时间段术语的文档词频保持在一个稳定水平，会出现较小幅度的增长甚至减少。

“云计算”这种热词的词频数量级高，成长趋势明显，短期内便可认定为新术语。大多数术语的数量级较低，但其词频成长分布依然存在以上趋势，以2014年《汉语主题词表》中冶金工业领域的已收录术语“连铸”一词为例，该术语在相应的1991年版《汉语主题词表》中尚未收录，其在该学科领域下文档词频随时间的分布情况及对应环比增长率如图1中(b)所示。1991年之前，即在1981－1990年间，“连铸”一词均保持每年有10篇左右的文献出现，其逐年环比增长率有较高值但却不满足后续的持续增长，即不满足“正向增长转折点”的条件，这一段时间即为新术语的“开始期”。该时期的特征是候选词的文档词频数在整个时间段内都处于一个较低水平，不曾出现增长率明显的“正向增长转折点”，而相应地，其在1991年《汉语主题词表》中未被选取；反而是在1993年，其词频增长率 $(42 - 22)/22 = 0.909$，且往后有超过5年以上的词频增长期，即1993－2004年度为该术语的成长期；而再往后，即从2005年往后，可认为这一时期属于该术语的饱和期。

同理，分别给出短期低频词“无线充电”、长期高频词“数字图书馆”的文档分布统计及对应环比增长率情况，如图1中(c)、(d)所示，其中“无线充电”的正向增长转折点确定为2009年，“数字图书馆”的正向增长转折点则确定为1995年。

可以看出，不同词频数量级的术语尽管由于研究热度及影响力的差异，其开始期的持续时间长短有所差异。比如高频词往往开始期持续时间极短，低频词的开始期则会持续多年，但它们都存在较明显的成长期，并经过成长期开始过渡到饱和期，即不同类别术语一般情况下都满足术语分布规律，而该分布规律可以有效反映术语的时间持续性及术语性特征，因此可用于判断识别新术语。本文从时间点及时间段角度出发，利用环比增长率指标，确定术语的两个关键时间点，出现时间点及正向增长转折点。出现时间点的确定可以有效确定对候选新术语观察时间段的确定，而术语自“正向增长转折点”后一般经过5年左右的时间可以达到与成熟期较接近的频数，为及时发现新术语，可认为当术语处于成长期一定年限，便可以判断其能作为成熟新术语，以添加到对应叙词表中。总之，将新术语的出现时间点 s 作为起始时间点，研究候选新术语的DF分布情况，当研究时间段内存在正向增长转折点，即增长趋势明显并持续保持一段时间，即认为该候选新术语处于成长期。当处于成长期的候选新术语超过“正向增长转折点”一定年限，则可以认为该术语同时具有新颖性、时间持续性及术语性，从而能添加到对应词表中作为新术语。

3 实证与分析

3.1 数据来源介绍

本文选择中国知网数据库作为候选词来源，选择该对象的原因是文献数据库关键词可以作为候选术语并用于叙词表编制及更新维护，该数据库中数据量足够大且覆盖时间范围较广，能获取本研究所需要的时间及学科领域信息。其中，时间指主题词对应文献的发表时间；学科领域信息则以中国知网数据库所采用的文献分类目录作为本文的学科领域分类；另外本文研究时间点选取为“每一年”。

3.2 新术语分布特征验证

为验证新术语一般性地处于成长期，即具有明显增长趋势，本文选取2014年出版的《汉语主题词表》[23]中矿业工程领域的一个词族“煤层”，其下位词术语有37个，见表3，其中有10个术语是相应的1991年出版《汉语主题词表》[24]中的已收录术语，记为实验组A。另外27个则是2014年《汉语主题词表》相比1991年的新增术语，记为实验组B。实证将分别统计以上已收录术语及新增术语文档词频的平均环比增长率，验证新增术语组B的环比增长率相比较已收录术语组A是否会有明显差别。

利用中国知网数据库高级检索功能，限定文献分类目录为“冶金工业”，在该学科领域下，限定1991－2017年的时间范围，以词族中各术语作为检索主题词，分别统计各年检出文献数，即文档词频值，得到对应三元组 (w, t_i, df)。以年为单位，分别统计各年A组和B组术语的平均文档词频值，其中，平均文档词频

表3 “煤层”词族术语

A组:1991年已收录术语		B组:2014年新增术语				
突出煤层	倾斜煤层	保护煤层	开采煤层	夹矸煤层	下部煤层	煤体结构
薄煤层	瓦斯煤层	本煤层	邻近煤层	坚硬煤层	卸压煤层	特殊煤层
多煤层	易燃煤层	单一煤层	破碎煤层	揭穿煤层	褶皱煤层	煤层特征
厚煤层	中厚煤层	断失煤层	浅埋煤层	深部煤层	自燃煤层	
缓倾斜煤层	过煤层	软底煤层	松软煤层	近距离煤层		
急倾斜煤层	三软煤层	稳定煤层	煤体构造	低透气性煤层		

值的统计以当年的实际词量作为除数。最后,在此基础上逐年计算A组及B组平均词频值的环比增长率,统计结果如表4所示:

表4 A、B两组逐年平均词频值及环比增长率统计

年份	A组平均词频	A组平均词频环比增长率(%)	B组平均词频	B组平均词频环比增长率(%)
1991	17.4		2.6	
1992	16.0	-8.05	2.1	-19.23
1993	16.3	1.88	3.5	64.84
1994	23.7	45.40	3.6	4.00
1995	20.6	-13.08	4.7	31.17
1996	20.5	-0.49	4.2	-11.06
1997	20.1	-1.95	4.2	-0.62
1998	21.8	8.46	5.2	23.96
1999	23.0	5.50	5.4	3.89
2000	28.7	24.78	5.2	-4.11
2001	29.7	3.48	4.8	-5.97
2002	35.5	19.53	7.0	43.65
2003	50.0	40.85	10.4	48.97
2004	58.5	17.00	12.1	16.79
2005	68.9	17.78	16.5	36.39
2006	71.5	3.77	18.3	10.76
2007	95.2	33.15	21.0	14.98
2008	110.3	15.86	26.5	26.06
2009	146.1	32.46	37.3	40.78
2010	167.2	14.44	40.5	8.43
2011	189.7	13.46	47.4	17.20
2012	222.0	17.03	52.6	10.85
2013	207.4	-6.58	57.3	8.87
2014	244.2	17.74	70.2	22.64
2015	234.9	-3.81	67.9	-3.32
2016	193.0	-17.84	60.3	-11.24
2017	191.9	-0.57	65.7	9.04

根据表4中A组和B组逐年的平均环比增长率值,制作折线图如图2所示。B组在2014年之前的绝大多数时间里,其环比增长率均高于同一时期的A组,说明新术语相比成熟的已收录术语在同一时期要有较为明显的增长趋势,增长幅度较高。相对应的已收录术语A组尽管其平均文档词频值要高,但数量相对稳定,增长较为平缓,说明已收录术语更多地处于术语的“饱和期”,而新术语则更多处于术语的“成长期”。另外,B组曲线的波动程度也明显高于A组,其反映了新术语出现的偶发性,不同的时间点会有不同的新术语出现,这也是其平均文档词频值较小的原因。总之,实证证明新术语更多地处于成长期,根据文档词频分布及其对应的环比增长率值可以量化术语的增长及分布趋势。当候选新术语保持正向增长趋势超过一定年限,则认为该术语同时具有新颖性、时间持续性及术语性特征,可以考虑将其更新到对应词表中。

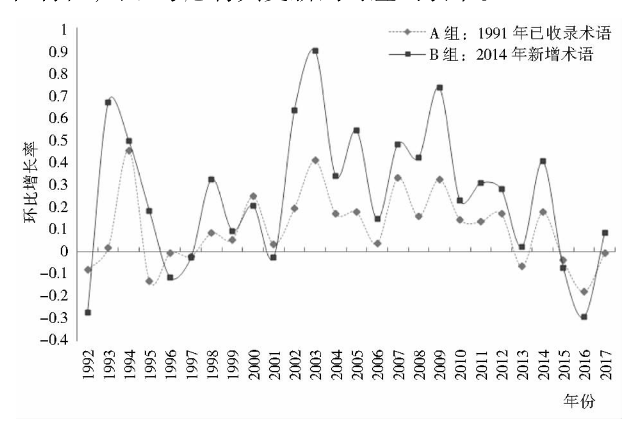


图2 A、B两组平均环比增长率对比情况

3.3 基于分布特征的新术语识别

以自动化技术及计算机技术领域(图书分类号为TP)为例,获取1989-2018年间,分类号为TP的全部文献元数据,共计获取256余万条。根据文献关键词在对应发表年份的文档词频值,共形成99余万条关键词时间序列。统计各关键词的总词频,过滤其中总词频小于20的关键词数据后,共计获取28 674个关键词及对应时间序列作为实验数据。确定各时间序列对应的开始时间点,并计算关键词各自对应的环比增长率序列,继而根据环比增长率值确定每个实验数据对应的“正向增长转折点”,不存在则记为0。

利用2014年最新出版的《汉语主题词表》[23]对关键词进行是否已收录的标注,以验证方法的有效性。确定“正向增长转折点”时,当阈值设为4和-0.5时,

即在“正向增长转折点”后连续 4 年 rg > －0.5 时，有 72% 的汉表已收录词，可以用本文方法预测到该词语收录到词表中的时间(预测收录时间 = 正向增长转折点时间 +4)。可见术语在成长分布中绝大多数都经历过术语成长期，而当阈值设为 4 以下及小于 －0.5 时，满足该规律的术语比例更大。但本文识别新术语时，为了保证识别的准确率，选择阈值为 4 和 －0.5。

对汉表未收录词，共计 10 296 个，使用同样方法，获取满足正向增长转折点后 4 年连续增长的数据，并过滤其中通用词，将预测收录时间为 2015－2018 年的关键词作为候选新术语，共计 673 个。经《汉语主题词表》编制相关专家判定，该部分 91.7% 的词均可以作为新术语补充到词表中。对候选新术语结果，按总词频排序，选取前、中、后各 10 个词示例如表 5 所示：

表 5　新术语识别部分结果展示

序号	侯选新术语				
1－10	深度学习	卷积神经网络	软件定义网络	CC2530	App
	网络舆情	安卓	Kinect	大数据时代	Android 系统
309－310	话题发现	混沌算法	局部敏感哈希	邻域信息	能量最小化
	社区检测	溯源系统	消息推送	虚拟机放置	短文本分类
665－674	图像模拟	蜕变关系	显著性图	消费行为	旋转森林
	遥感(RS)	有源 RFID	语义相关性	预警信息	元启发式算法

本文方法在新术语收录判断中具有较高的准确率，分析候选新术语对应的总词频，不仅高频新术语可以被识别出来，例如“深度学习”(df =1603)、“卷积神经网络”(df =1396)等词，同时实际应用中占比较多的低频新术语也能被识别出来，例如“元启发式算法”(df =20)、“有源 RFID”(df =20)等。

4　结论与展望

本文基于新术语的新颖性和术语性特征，以《汉语主题词表》已收录术语为例，通过相关统计分析，研究术语的时间分布特性和文档词频分布特性。从时间点及时间段角度出发，利用环比增长率指标，确定术语的两个关键时间点，出现时间点及正向增长转折点，出现时间点作为观察术语成长分布的起点，正向增长转折点则作为划分术语开始期及成长期的标志时间点，并利用文档词频逐年分布的趋势图，研究术语在开始期、成长期以及饱和期的文档词频分布差异以及指标变化情况。实证证明新术语更多地处于成长期，当候选新术语保持正向增长趋势超过一定年限，则认为该术语同时具有新颖性、时间持续性及术语性特征，可以考虑将其更新到对应词表中。通过对自动化技术及计算机技术领域的新术语识别效果分析，证明本文提出的新术语分布规律可以有效应用于叙词表的新术语收录评估中。

由于目前的研究所采用的学科领域是指已经确定好的领域，而当候选新术语属于新兴领域及小规模领域时，新术语的成长分布情况可能会有所差别。因此，还需要在更多学科领域中进一步完善和发展本文的研究。

参考文献：

[1] 冷伏海，徐跃权，冯璐. 信息组织概论：第 2 版[M]. 北京：科学出版社，2008：197.
[2] 周晓英，曾建勋. 主题词表的社会应用研究[J]. 数字图书馆论坛，2014(10)：2－6.
[3] 常春. 网络环境下叙词表编制与发展[M]. 北京：科学技术文献出版社，2015：101－103.
[4] 中国科学技术信息研究所.《汉语主题词表》服务系统[EB/OL]. [2018－11－20]. https://ct. istic. ac. cn/site/organize/word.
[5] 侯丽，李姣，侯震，等. 基于混合策略的公众健康领域新词识别方法研究[J]. 图书情报工作, 2015,59(23):115－123.
[6] 荀恩东，李晟. 采用术语定义模式和多特征的新术语及定义识别方法[J]. 计算机研究与发展，2009，46(1)：62－68.
[7] 邢恩军，赵富强. 基于上下文词频词汇量指标的新词发现方法[J]. 计算机应用与软件，2016，33(6)：64－67.
[8] 刘辉，刘耀. 基于条件随机场的专利术语抽取[J]. 数字图书馆论坛，2014(12)：46－49.
[9] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International journal on digital libraries, 2000, 3(2): 115－130.
[10] 熊李艳，谭龙，钟茂生. 基于有效词频的改进 C-value 自动术语抽取方法[J]. 现代图书情报技术，2013，29(9)：54－59.
[11] 胡阿沛，张静，刘俊丽. 基于改进 C-value 方法的中文术语抽取[J]. 现代图书情报技术，2013，29(2)：24－29.
[12] 韩红旗，安小米. C-value 值和 Unithood 指标结合的中文科技术语抽取[J]. 图书情报工作，2012，56(19)：85－89.
[13] WANG M, LIN L, WANG F. New word identification in social net-

work text based on time series information[C]//IEEE. International conference on computer supported cooperative work in design. New York:IEEE,2014:552 – 557.
[14]黄轩,李熔烽. 博客语料的新词发现方法[J]. 现代电子技术,2013,36(2): 144 – 146.
[15] 邹纲,刘洋,刘群,等. 面向 Internet 的中文新词语检测[J]. 中文信息学报,2004, 18(6):1 – 9.
[16] 吴悦,燕鹏举,翟鲁峰. 基于二元背景模型的新词发现[J]. 清华大学学报(自然科学版), 2011, 51(9):1317 – 1320.
[17] 苏其龙. 微博新词发现研究[D]. 哈尔滨:哈尔滨工业大学,2013:43.
[18] LIU W, SU J, LEI X, et al. Graph-based equivalence concept matching in knowledge organization system integration: a case study on thesaurus [C]//IEEE. International conference on natural computation, fuzzy systems and knowledge discovery. New York: IEEE,2018:839 – 844.
[19] 高永伟. 近 20 年英语国家对新词的研究[J]. 外语与外语教学,1998(11):9 – 11.
[20] 亢世勇. 新词语大词典[M]. 上海:上海辞书出版社,2003.
[21] 刘长征. 新词语的生命力[J]. 北华大学学报(社会科学版),2012,13(5):4 – 8.
[22] 常春,杨婧. 基于生物种群增长规律的概念词频变化特征研究[J]. 情报科学, 2018,36(10): 128 – 132.
[23] 中国科学技术信息研究所. 汉语主题词表:工程技术卷(第二分册)[M]. 北京:科学技术文献出版社, 2014:425 – 426.
[24] 中国科学技术信息研究所. 汉语主题词表:自然科学增订版(第三分册)[M]. 北京:科学技术文献出版社, 1991:353.

作者贡献说明:
雷晓:提出思路,完成论文撰写;
常春:参与论文设计,修改论文;
刘伟:参与论文设计,修改论文。

Research on the Distribution Characteristics of New Terminology for the Update of the Thesaurus

Lei Xiao　Chang Chun　Liu Wei
Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] In order to enhance the practicability of thesaurus, it is necessary to constantly update new terms in the field to thesaurus. In the process of updating and maintenance, we should explore the distribution characteristics of new terms from the perspective of time and frequency, which can provide reference for the method of discovering new terms. [**Method/process**] Based on the relevant characteristics of the new terminology, combined with the development distribution of the corresponding document frequency at time point and period, through the relevant statistical analysis, the distribution of terminologies in different development periods is studied, especially the characteristics of terminologies from the beginning to the maturity. [**Result/conclusion**] It is proved that the new terminology is generally in the growth stage of terminology. When the candidate new terminology keeps positive growth trend for more than a certain number of years, it is considered that the term has all novelty, time persistence and terminological features. Based on the distribution characteristics, the article selects a subject area to discover its new terminology. According to the judgment of the expert, the method has a high accuracy rate in the judgment of new term, and can effectively identify the low frequency words which are more occupied in practical applications.

Keywords: thesaurus update　new terminology　time distribution　document frequency distribution